Analyzing XploRe Download Profiles with Intelligent Miner

Hizir Sofyan and Axel Werwatz[1]

[1] Institute für Statistik und Ökonometrie, Humboldt Universität zu Berlin, Spandauer Str. 1, 10178, Berlin

Summary

This paper is an example of data mining in action. The database we are mining contains 1085 profiles of individuals who have downloaded the statistical software `XploRe`. Each profile contains the responses to an online questionnaire comprised of questions about such things as an individuals' computing preferences (operating system, favourite statistics software) or professional affiliation. After formatting and cleaning the raw data using `MS Excel`, we use `IBM`'s `Intelligent Miner` to perform a cluster analysis of the download profiles.

We try to identify a small number of "types" of users by employing a clustering algorithm based on the *New Condorcet Criterion*, which is particularly well-suited for our all-categorical data. We identify three clusters in the mining run: `Academia`, `Unix/Linux users` and `Researchers`. The three variabels that are most important in identifying the clusters are an individual's kind of work, the way he or she got to know `XploRe` and the operating system of his or her computer.

1 Introduction

Recently, the capability to both generate and collect data has been expanded enormously and provides us with huge amounts of data that are often routinely collected during daily operations. To store, organize and access the data powerful and affordable database management systems are available (Ha, 1998).

The aim of data mining is to -intelligently and automatically- extract useful information from these databases. It tries to discover patterns and relationships hidden in the data using suitable statistical models and techniques. Thus, data mining may yield profitable results for almost every organization that collects data on its customers, markets, products or processes.

This paper is an example of data mining in action. The database we are mining contains the profiles of individuals who have downloaded the statistical software `XploRe`. `XploRe` is aimed at sophisticated users who are looking for a flexible, programmable statistics package with an emphasis on more advanced procedures (Härdle, 1999). It is therefore particularly important for the makers of `XploRe` to get to know those who are interested in their product.

Before a free trial version of `XploRe` can be downloaded , customers are asked in an online questionnaire to provide information, for instance, about their identity, occupation and computing preferences. The raw data thus consists of 1085 multivariate profiles of individuals who downloaded `XploRe` between November 1, 1999 and October 31, 2000.

Our mining run consists of the following steps:

- organize and clean the ASCII raw data with `MS Excel`
- transfer the data to IBM's `Intelligent Miner`
- perform a cluster analysis of the download profiles

In the cluster analysis, we try to identify a small number of "types" of users. We identify three clusters: `Academia, Unix/Linux users` and `Researchers`. The three variabels that are most important in identifying the clusters are an individual's kind of work, the way she got to know `XploRe` and the operating system (`Windows`, `Unix`,..) for which she downloaded `XploRe.`

The remainder of the paper is organized as follows. In the next section, we describe the online questionnaire used to collect the data and how the raw data was organized and cleaned. Before we turn to the cluster analysis, we summarize the data and explain our choice of mining software and clustering algorithm. This is followed by a discussion of the results of the cluster analysis. The final section concludes.

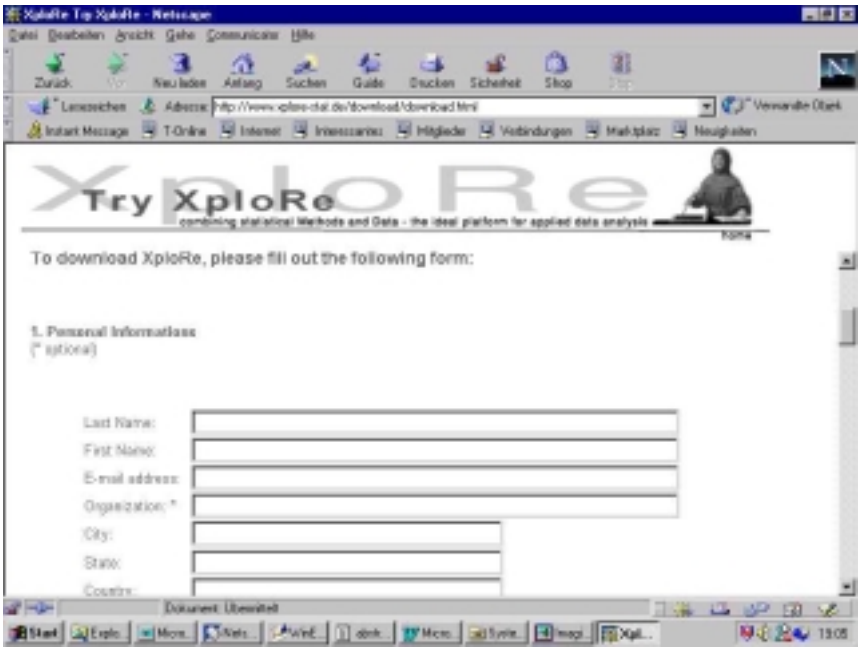

Figure 1: **Personal Questions**

2 The data

2.1 Getting the data

Those interested in the statistical computing environment `XploRe` can download a free trial version from its homepage. All versions of `XploRe` (except for the Linux local version) are demo-versions that expire after about two months, are limited to 1000 observations and do not include all of `XploRe`'s features and commands. The Linux local version has no expiration date and no limit on the size of the data.

Before the actual download is started, customers are asked to interactively provide information about their identity, computing preferences, etc.

The first part of the online questionnaire is shown in Figure 1. Customers are prompted to type in their name, affiliation and address. All of these items, except for the affiliation, must be answered to be able to download `XploRe`. We will refer to this group of questions as "personal questions".

The next set of questions is shown in Figure 2. Users are asked to provide information on the `kind of work` they need `XploRe` for, the way they learned about `XploRe` (`get information`) and the `statistical software` they currently use. All these questions are answered by selecting items from a list of possible responses. Answering them is not required for downloading

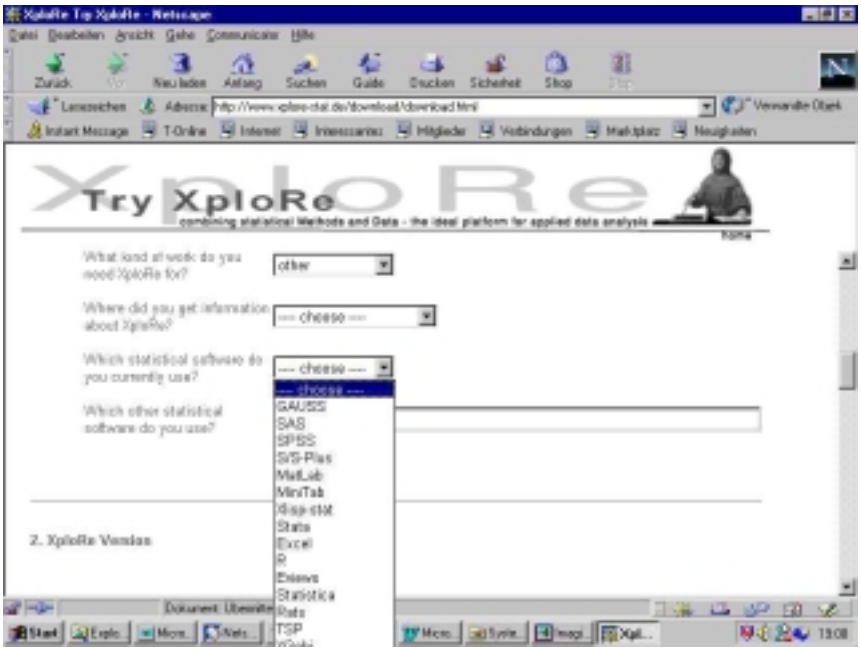

Figure 2: **Substantive Questions**

`XploRe` but their optional character is not explicitly revealed. The responses to these questions will play a prominent role when we try to form groups of homogeneous users. This is the reason for referring to them as "substantive questions".

The last part of the questionnaire, not shown here to save space, contains rather "technical" questions. Users have to select the version of `XploRe` they would like to download. They can choose from a list of `operating systems` (`Windows, Linux`, etc.) and between the local version of `XploRe` and the Java-Client-only version. They also can opt to subscribe to the `XploRe` mailing list.

As part of the download procedure, the `date` and `IP-address` are automatically recorded. The raw data thus consists of multivariate profiles for 1085 individuals who downloaded `XploRe` from November 1, 1999 to October 31, 2000.

2.2 Cleaning the data

Even the most sophisticated statistical methods will deliver erroneous results if the data are of poor quality or simply wrong: "garbage in, garbage out", as the saying goes. We therefore thoroughly cleaned the raw data before using it in the cluster analysis.

Those that download XploRe clearly would like to complete the downloading process as swiftly as possible and probably prefer to be asked no questions at all. If asked too many, too complicated, or too curious questions people might get annoyed and give wrong or incomplete answers.

In the case of XploRe, there are only a few "substantive" questions and, as shown in Figure 2, they are all easy to answer. But they are preceeded by the mandatory questions on a user's identity to which quite a few persons gave an intentionally wrong answer. We took obviously false answers to the "personal questions" as indicators that the answers to the "substantive questions" may also be wrongly answered. Suspicious observations were inspected one-by-one and dropped according to a set of criteria.

It should be pointed out, that this part of the data mining process can hardly be made fully automatic. There are simply more ways to supply false information than any computer program can identify.

2.3 Summary of the data

Before turning to the cluster analysis, we will give a summary of the variables that comprise the download profiles.

All of our variables are categorical (which will be reflected in our choice of clustering method). Hence, it makes sense to use the modal categories of each variable to create the "typical" person who downloaded XploRe. This person works at an European university, has heard of XploRe on the web, uses MS Excel to do statistics, and has downloaded the local Windows version of XploRe.

Turning to the variables one-by-one, about 80% of the users work either at a university (46%) or a research institution (34%) while only a small fraction of the downloads are initiated by people working for private, non-research companies.

Figure 3 shows that information about XploRe has been disseminated through various channels, with the world wide web already playing the most important role.

Two-thirds of all downloads request the Windows version of XploRe, with Linux coming in second at about 30%. The popularity of the Linux version is surely boosted by its relative attractiveness (no expiration date, no limit on size of datasets). All other operating systems play a very minor role.

We were quite surprised to learn that when people were asked about the statistical software they currently use, the most frequent response was MS Excel. XploRe is a command driven software with a comparative advantage in rather advanced statistical methods, particularly smoothing. We expected programs similar in design and scope to XploRe such as S-Plus or GAUSS to be the most frequent choices in this category. While there are substantial shares of S-Plus and GAUSS users (both account for about 7% of all downloads), Figure 4 shows that most downloads are made by users of more standard (statistics) software such as MS Excel, SPSS or SAS. Some of these users

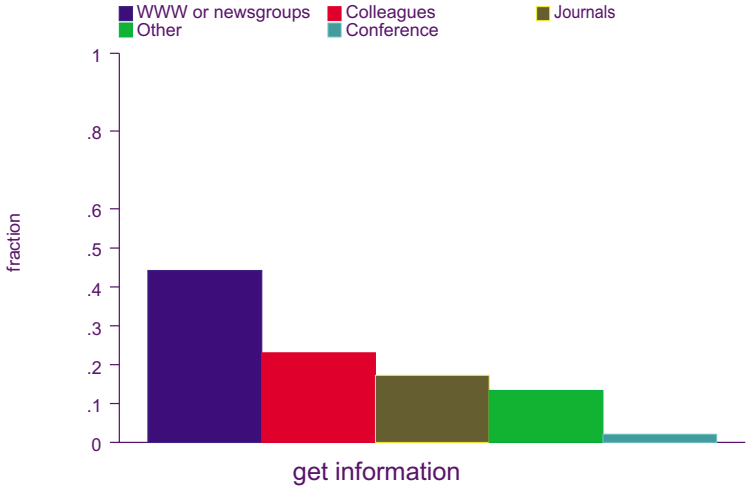

Figure 3: **Get Information**

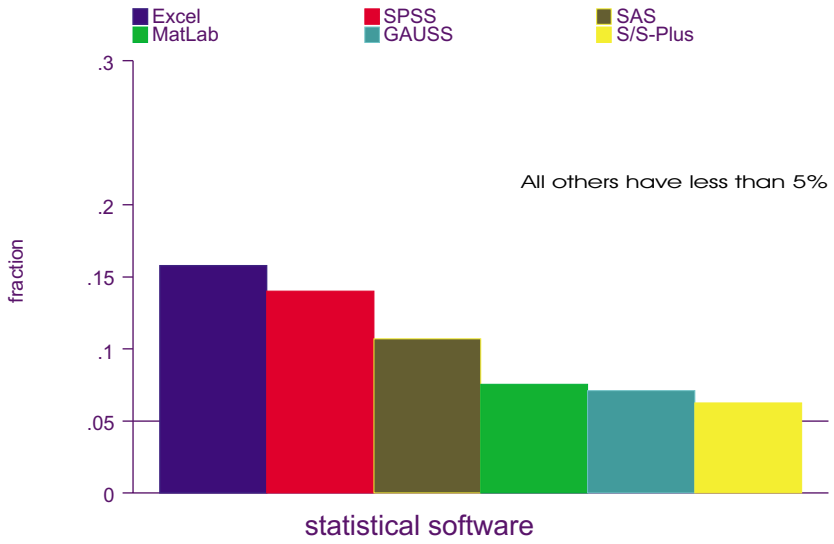

Figure 4: **Statistical Software**

may be facing a statistical problem that requires a programmable, matrix-oriented tool like `XploRe`.

Finally, the majority of the downloads are made by residents of Germany (22.5%), the U.S. (almost 18.8%), and Italy, France, and the UK (which combine for 17.51%). About 10% of the downloads are made from Asia while Australian and African downloads are very rare.

3 Mining Technology

3.1 Mining Software

The primary goal of our mining run is to identify a small number of "types" of users by performing a cluster analysis of the download profiles. We chose `Intelligent Miner` to do the cluster analysis for several reasons.

Firstly, the clustering algorithm of `Intelligent Miner` employs the *New Condorcet Criterion* (NCC), which is particularly well-suited for a data set like ours that consists entirely of discrete (categorical) variables. We will discuss NCC in more detail in the following section.

Secondly, `Intelligent Miner` comes with a cluster visualization tool that greatly helps in solving one of the main problems of cluster analysis: figuring out what the clusters mean and translating those insights into meaningful customer segments. An example is shown in Figure 5 of section 4 where we will explain and utilize `Intelligent Miner`'s graphical output.

Finally, the well tested parallel algorithm of `Intelligent Miner` is computationally efficient.

3.2 Clustering Method

Cluster analysis partitions the observations into disjoint groups (synonyms are "classes" and "clusters") such that observations belonging to the same class are very similar while observations belonging to different classes are very different. The similarity of observations is measured by some **distance** function (such as the Euclidean distance for continuous variables). Once a rule for computing the distance between any two observations has been determined, one can compute for a given partition the distances between observations of the same class (within-group or *intra*class distances) and observations of different classes (between-group or *inter*class distances). These individual distances are combined by a partition **criterion** that can be used to compare different partitions. Since it is in general impossible to consider all possible partitions of the observations, a clustering algorithm also has to have a **strategy** for searching for the optimal partition.

Various choices exist for all aspects of this problem: the distance, the partition criterion and the optimization strategy. We will not attempt to provide a comprehensive discussion here but rather focus on the algorithm

we have used and point out its advantages.

Condorcet Criterion
The clustering algorithm used in this paper is based on the New Condorcet Criterion (NCC) of Michaud (1997). It is inspired by Condorcet (1743-1794)'s work on finding a desirable way to aggregate votes (rankings) in an election. The NCC combines intraclass agreement as well as interclass disagreement such that "good" partitions, i.e. those with small intraclass distances and large interclass distances, get higher values of the criterion function.

Specifically, for a given partition P, the goodness criterion is

$$G(P) = \sum_{k=1}^{p} \sum_{i \in L_k} \left(\sum_{j \in L_k; j \neq i} (m - d_{ij}) + \sum_{j \notin L_k} d_{ij} \right). \quad (1)$$

The index of the leftmost sum is over the p classes (clusters) while the inner sums run over the observations that are within and outside some cluster L_k. The summands are the distances d_{ij}.

For categorical variables, the distance d between two observations i and j is the number of variables for which the two observations take on different values, i.e. the number of *disagreements* between the two observations. If m variables are measured for each observation then it follows that $m - d_{ij}$ is just the opposite of d_{ij} : the number of *agreements* between observations i and j.

$G(P)$ calculates for a given partition P the sum of all intracluster agreements and all intercluster disagreements. Different partitions may then be ranked according to their value of $G(P)$ – the higher $G(P)$, the better the partition. Indeed, if we interpret intracluster agreements and intercluster disagreements as "votes" for a given partition, then the connection to Condorcet's work becomes apparent: the winner among all candidates (i.e. partitions) of the election (i.e. the cluster analysis) is the one receiving the most votes (i.e. the highest value of $G(P)$).

It remains to explain `Intelligent Miner`'s strategy for finding the optimal partition (i.e. the one with the highest value of $G(P)$). `Intelligent Miner` considers all partitions whose number of clusters p is smaller than the maximum number of clusters specified by the user. That is, if the user wants to find the best partition among those with, say, at most three clusters then `Intelligent Miner` will calculate $G(P)$ for all partitions with one, two or three clusters.

4 Results and Discussions

The clustering algorithm described in the previous section requires two inputs from the user: the variables to be included in the analysis and the

maximum number of clusters per partition. We have tried various combinations of variables and maximum number of clusters, basically following a heuristic "backward selection" strategy. That is, we started with a relatively large number of variables and maximum number of clusters, aiming to find a smaller, more parsimonious partition.

In this paper, we will show two stages along this path. First, we present the results of a partition based on five variables and a maximum of five clusters per partition. Then we present a more parsimonious, "optimal" partition, based on three variables and a maximum number of three cluster.

N=1085 Global Condorcet Value = 0.4722			
Id	Cluster Size (Absolut)	Cluster Size (Relative)	Condorcet Value
0	324	29.86	0.4923
1	220	20.28	0.5121
2	191	17.60	0.4584
3	203	18.71	0.4402
4	147	13.55	0.3690

Table 1: Cluster Characteristics

Tables 1, 2 and 3 summarize the results of the cluster analysis with five variables and a maximum of five clusters. The five "active" variable (that were actually used in forming the clusters) are `kind of work`, `statistical software`, `get information`, `country`, and `operating system`. There are also two supplemental variables (`get on mailing list`, `kind of server`).

Similarity Filter : 0.30		
Cluster 1	Cluster 2	Similarity
1	4	0.30
2	3	0.31

Table 2: Similarity Between Clusters

Table 1 gives the value of the overall goodness criterion $G(P)$ of equation (1) ("Global Condorcet Value = 0.4722"), as well as the *intra*cluster agreements (column "Condorcet Value") for each of the five clusters. Table 2 reports *inter*cluster agreements for selected clusters. All measures are standardized to fall between 0 and 1. A perfect partition would attain a value of

1 for the global condorcet value and each *intra*cluster agreement and a value of 0 for each *inter*cluster agreement.

The global condorcet value and all *intra*cluster agreements obviously leave room for improvement. The reported *inter*cluster similarities between clusters 1 and 4 (0.30) and clusters 2 and 3 (0.31) are both relatively high. This suggests to merge these pairs of clusters to arrive at a more parsimonious partition.

Id	Name	Modal Value	Modal Freq.	No of Values	Condorcet Value
1	OS Platform	5	65.53	5	0.5342
2	Kind of Work	1	45.90	5	0.3488
3	Get Info	2	44.15	6	0.2960
4	Software Stat.	Excel	15.76	16	0.1004
5	Country	Germany	22.49	68	0.1026
6	[Mail List]	0	80.65	2	-
7	[Server]	1	96.77	2	-

Table 3: Variable Characteristics

Regarding the role of the variables in forming the clusters, Table 3 reports –additional to some summary statistics– a χ^2-type value (also called "Condorcet Value") that reflects how much the distribution of each variable varies across clusters. The distribution of the operating system (`OS Platform`) shows the greatest variation across the five clusters. Hence, this variable played the greatest role in seperating the observations into the observed partition. `kind of work` and `get info` are the only other variables with condorcet value of more than 0.30. We therefore chose to continue the analysis with these three variables only and do not give an interpretation of the five-cluster partition.

N=1085 Global Condorcet Value = 0.6166			
Id	Cluster Size (Absolut)	Cluster Size (Relative)	Condorcet Value
0	420	38.71	0.6355
1	352	32.44	0.5977
2	313	28.85	0.6063

Table 4: Cluster Characteristics with three clusters

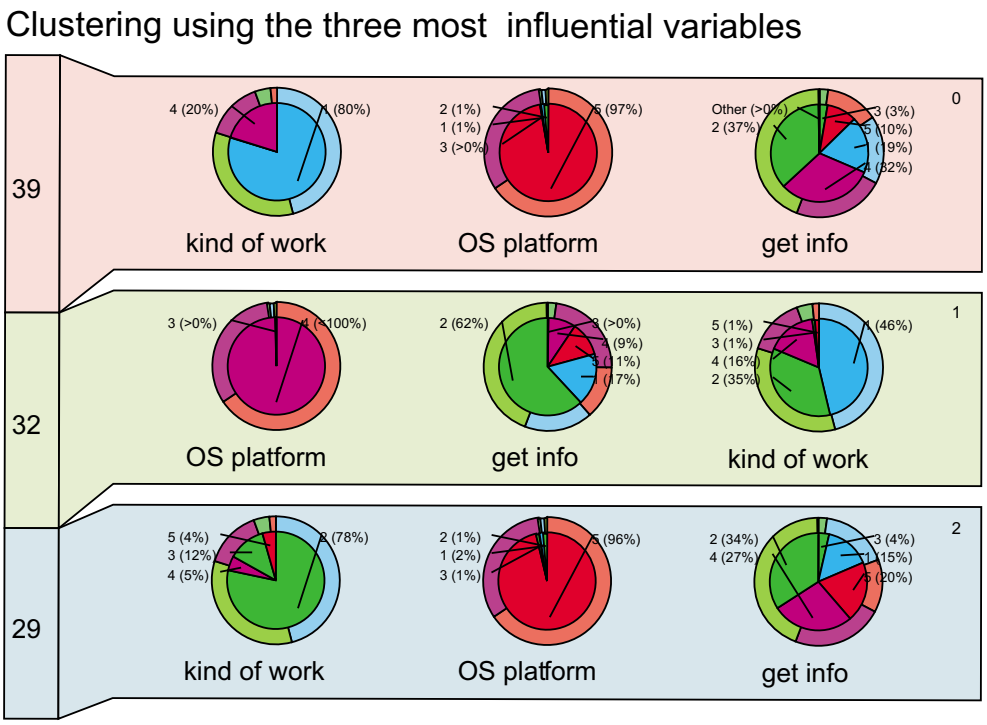

Figure 5: **Three clusters (cluster I (39% of all observations) ; cluster II (32%) ; cluster III 29%**

The results of the cluster analysis with three variables and at most three clusters per partition are presented in Tables 4 and 5 and Figure 5. The global condorcet value (see Table 4) for the optimal partition in this setting is 0.62 which is a substantial improvement from the partition of Table 1. Note that the condorcet values of the *intra*class agreements all exceed 0.5.

The interpretation of the clusters is made easier by the output of `Intelligent Miner`'s cluster visualization tool shown in Figure 5. There are three rows of graphs, one row for each cluster. Each row has pie charts for each variable, ordered from left to right according to decreasing influence in forming the clusters. Each pie charts presents the distribution of the respective variable among the observations in the cluster (inner pie) and the distribution among all observations in the entire sample (outer pie).

For instance, the leftmost pie chart in the first row of Figure 5 shows the distribution of the variable `kind of work` among the observations in the first cluster (inner pie) and among all observations (outer ring). In the first cluster (which contains 39% of all observations), 80% of the profiles have a value of "1" (university) for `kind of work` (turquoise slice), while the remaining 20%

of the members of cluster I have a value of "4" (private; purple slice). From the outer ring of the pie chart, we can see that the corresponding fractions (represented by turquoise and purple segments of the outer ring) among *all* observations of the dataset are considerably smaller (46% download `XploRe` for work at the university and 14.5% for private use).

In the second row (cluster II), the pie chart for `kind of work` is the rightmost pie chart, indicating that `kind of work` is the least influential variable for allocating observations to cluster II. If we compare the inner pie with the outer ring of this pie chart, it is easy to say why: the distribution of `kind of work` in cluster II and in the entire sample are very similar.

By studying all pie charts in this way, one can arrive at the charaterization of the three clusters summarized in Table 5.

Type Cluster	Variables	Attributes	Percentage
Cluster I (`Academia`)	`kind of work`	university	80%
	`OS platform`	`Windows 98/NT`	97%
	`get info`	publication, journal	19%
		friends, colleagues	32%
		WWW, newsgroups	37%
Cluster II (`Unix/Linux users`)	`OS platform`	`Unix/Linux`	99%
	`get Info`	WWW, newsgroups	62%
	`kind of work`	university	46%
Cluster III (`Researchers`)	`kind of work`	research	78%
	`OS platform`	`Windows 98/NT`	96%
	`get info`	others	20%
		friends, colleagues	27%
		WWW, newsgroups	34%

Table 5: Three clusters

Clusters I and III are best characterized by their dominant value of `kind of work` (university and research, respectively), leading to our choice of names for these clusters (`Academia` and `Researchers`, respectively). Regarding the other two variables, `OS platform` and `get info`, the observations in both clusters are quite similar: `Windows` is the dominant opertaing system and the world wide web the most prevalent source of information about `XploRe`.

Cluster II, on the other hand, is best characterized by `OS platform`: 99%

of the individuals in this cluster downloaded the `Linux` or `Unix` version of `XploRe`. Compared to the `Academia` and `Researchers` clusters, the World Wide Web plays an even more important role as a source of information: 62% of the members of cluster II learned about `XploRe` from the web. Regarding `kind of work`, we have already noted above that members of the `Linux/Unix` cluster closely mimic the distribution among all observations. Hence, usage of `Linux/Unix` usage appears to be independent from `kind of work` in the statistical sense.

Since installing and using software under `Linux` or `Unix` still requires more sophistication on the part of the user, one may conclude that this cluster is made up of "computer experts", who also tend to interact with others "electronically". Members of clusters I and III, to the contrary, overwhelmingly use some (typically preinstalled) version of `Windows`. While the latter also increasingly rely on the web for information purposes, they still also use more traditional ways of communicating with their peers, such as scientific publications or conferences.[1]

From a business perspective, these clusters suggest some strategies for promoting `XploRe`. Members of the `University` cluster will use `Xplore` for research *and* teaching. Hence, members of this group (some of which are likely to be students) will be interested in `XploRe`'s capabilities to teach and learn statistics, such as its "teachware" modules.

Members of the third cluster, on the other hand, will primarily want to know how `XploRe` can help them with their research work. Adding questions to the online questionnaire about the field of specialization (economics, epidemeology, etc.), the statistical techniques typically used or the statistical methods searched for in `XploRe` may provide valuable additional information for targeted marketing and the future direction of `XploRe`.

Regarding a marketing strategy for the members of the `Linux` cluster, they appear to be a target group for promoting `XploRe`'s web-based features, such as the online help system or its ability to do all computations "through" an internet browser.

Their reponses (as well as, to a lesser degree, those of the members of the other clusters) also show the importance for a statistical software like `XploRe` to have an excellent WWW appearance and to know how to get the "online" attention of its (potential) customers. That is, the makers of `XploRe` (and most likely those of any other sophisticated statistical software) have to make sure that people will find and keep visiting the `XploRe` homepage and read (and not immediately discard) `XploRe` email.

[1] Indeed, in an earlier analysis with less recent profiles, the world wide web was less important in clusters I and III. The increased importance is probably due to both the general increase in internet usage, as well as the enhanced internet representation of `XploRe`.

5 Summary and Conclusions

In what has been called the "information age" or the "digital era", everyday economic activity continuously produces large amounts of data, stored in ever growing data bases. The goal of "data mining" is to -intelligently and automatically- extract useful information from these databases.

One of the main tasks of data mining is clustering, i.e. partitioning the data into (a small set of) groups such that members of the same group are rather similar while members of different groups are rather different. Such a partition may help in producing a company to design differentiated promotion and sales strategies aimed at certain "types" of customers.

In this paper, we have presented the results of a cluster analysis of 1085 profiles of individuals who have downloaded the statistical software `XploRe`. Each profile consisted of a set of variables that are the responses to a mandatory online questionnaire preceeding the actual downloading process.

Using a clustering algorithm particularly suited for our categorical data, we arrived at a partition consisting of three clusters: `Academia`, `Linux/Unix users` and `Researchers`. The first and third cluster both consist of persons who work under `Windows` and learned about `XploRe` in various ways, with the world wide web playing an increasing but not (yet) dominant role. These two clusters are separated primarily by the `kind of work` variable: members of the `Academia` cluster predominantly work at universities and will use `XploRe` for research *and* teaching. Members of the second cluster can be characterized as sophisticated computer users who work under `Linux` or `Unix` and make heavy use of the internet. For each cluster, we have tried to sketch a marketing startegy that incorporates the results of this analysis.

From a statistical point of view, two possible improvements always come to the analyst's mind: better data and better methods.

Indeed, we would have liked to have a greater number of substantive variables at our disposal (for instance, the field of specialization or the most frequently used statistical methods). On the other hand, the extensive data cleaning job showed that a longer online questionnaire need not necessarily lead to better data (as well as that data mining, at least in this analysis, is not an all-automatic affair).

Regarding the methods, one may experiment both with other data mining techniques (such as association rule or sequential mining) or other clustering algorithms to find new or improved results. We leave this for future work.

References

Chen, M. S., Han, J. & Yu. P. S. (1996). *Data Mining: an Overview from a Database Perspective*, IEEE Trans. on Knowledge and Data Engineering, 8:866-883.

Grabmeier, J. & Rudolph, A. (1998). *Techniques of Cluster Algorithms in Data Mining*,IBM.

Ha, S. H. & Park, S. C. (1998). *Application of data mining tools to hotel data mart on the Intranet for database marketing*, Expert System with Application 15:1-31.

Härdle, W. & Klinke, S.& Müller, M. (1999). *XploRe Learning Guide*, Springer Verlag, Hiedelberg.

Härdle, W. , Hlavka, Z. & Klinke, S. (2000). *XploRe Application Guide*, Springer Verlag, Hiedelberg.

Michaud, P. (1987). *Applied Stochastic Models and Data Analysis*, 3:173-189.

Michaud, P. (1997). Clustering Techniques, *Future Generation Computer Systems*, 13:135-147.